Evaluation of Quantitative Structure−Activity Relationship Methods for Large-Scale Prediction of Chemicals Binding to the Estrogen Receptor[†]

Weida Tong,*[,‡] David R. Lowis,[§] Roger Perkins,[‡] Yu Chen,[‡] William J. Welsh,[△]
Dean W. Goddette,[§] Trevor W. Heritage,[§] and Daniel M. Sheehan[¶]

R.O.W. Sciences, Inc., Jefferson, Arkansas 72079, Department of Chemistry & Center for
Molecular Electronics, University of Missouri-St. Louis, St. Louis, Missouri 63121, Tripos Inc.,
St. Louis, Missouri 63144, and Division of Genetic and Reproductive Toxicology, National Center for
Toxicological Research (NCTR), Jefferson, Arkansas 72079

Received January 27, 1998

Three different QSAR methods, Comparative Molecular Field AnaIysis (CoMFA), classical QSAR (utilizing the CODESSA program), and Hologram QSAR (HQSAR), are compared in terms of their potential for screening large data sets of chemicals as endocrine disrupting compounds (EDCs). While CoMFA and CODESSA (*Co*mprehensive *De*scriptors for *S*tructural and *S*tatistical *A*nalysis) have been commercially available for some time, HQSAR is a novel QSAR technique. HQSAR attempts to correlate molecular structure with biological activity for a series of compounds using molecular holograms constructed from counts of sub-structural molecular fragments. In addition to using r^2 and q^2 (cross-validated r^2) in assessing the statistical quality of QSAR models, another statistical parameter was defined to be the ratio of the standard error to the activity range. The statistical quality of the QSAR models constructed using CoMFA and HQSAR techniques were comparable and were generally better than those produced with CODESSA. It is notable that only 2D-connectivity, bond and elemental atom-type information were considered in building HQSAR models. Since HQSAR requires no conformational analysis or structural alignment, it is straightforward to use and lends itself readily to the rapid screening of large numbers of compounds. Among the QSAR methods considered, HQSAR appears to offer many attractive features, such as speed, reproducibility and ease of use, which portend its utility for prioritizing large numbers of potential EDCs for subsequent toxicological testing and risk assessment.

INTRODUCTION

The possibility that certain man-made chemicals can disrupt the sensitive endocrine systems of humans and other vertebrates by mimicking endogenous hormones has sparked intense scientific discussion and debate in recent years.[1] This growing national concern has resulted in legislation, including reauthorization of the Safe Drinking Water Act and passage of the 1996 Food Quality Protection Act, mandating that the Environmental Protection Agency (EPA) develop a screening and testing program for endocrine disrupting compounds (EDCs).[2,3]

The EDC issue and the pressing regulatory requirements portend a prodigious financial burden for screening and testing that will likely comprise a suite of *in vitro* and *in vivo* assays for multiple endpoints. Some 80 000 or more existing chemicals, many commercially important and produced in enormous quantities, may ultimately need to be evaluated for their estrogenic activity under the EPA mandate.[4] With the advent of combinatorial synthesis[5] and high-throughput screening[6] techniques, the number of chemicals to be tested is expected to grow dramatically in the coming years. Fortunately, this challenge is offset by the ability to construct quantitative structure−activity relationship (QSAR) models for the rapid prediction of activity. Such models have great potential for use in the identification and classification of large numbers of potential EDCs. At the very least, QSAR models could be employed to establish a prioritization procedure for subsequent biological testing.

An EDC can be broadly defined as “an exogenous agent that interferes with the production, release, transport, metabolism, binding, action or elimination of natural hormones in the body responsible for the maintenance of homeostasis and the regulation of developmental processes”.[1] Of the many biological mechanisms that can result in endocrine disruption, by far the most dominant and well studied is expression of an estrogenic response.[7,8] Although several mechanistic events can determine the *in vivo* estrogenic potency of a chemical, expression of an estrogenic response generally requires binding to the estrogen receptor (ER).

In recent years, several QSAR models have been developed for estrogenic compounds binding to the ER.[9−14] Most of these studies have employed the three-dimensional (3D)-QSAR method of comparative molecular field analysis (CoMFA)[15] for model building. This method requires a procedure known as “structural alignment” of the molecules under study because a common binding site is assumed. The utility of CoMFA has been demonstrated in a wide range of

[*] To whom all correspondence should be addressed. E-mail wtong@nctr.fda.gov.

[†] The opinions expressed are those of the authors and not necessarily those of the U.S. Food and Drug Administration.

[‡] R.O.W. Sciences, Inc.

[§] Tripos Inc.

[△] Department of Chemistry.

[¶] Division of Genetic and Reproductive Toxicology.

S0095-2338(98)00008-0 CCC: $15.00
Published on Web 05/20/1998

applications.[16−18] However, CoMFA requires some knowledge or hypothesis regarding the functionally active conformations of the molecules under study as a prerequisite for structural alignment. Moreover, care must be exercised when constructing molecular alignments because slight differences in alignment can lead to wide variations in the resultant CoMFA model.

Classical QSAR models were also considered in the present study, and were produced using partial least-squares (PLS) multivariate linear regression techniques. Classical QSAR techniques attempt to correlate a biological activity or a physical property of interest with a pre-defined set of calculated physicochemical descriptors within a collection of related compounds. In contrast to CoMFA, classical QSAR methods require no structural alignment of the molecules.[9] However, both CoMFA and classical QSAR methods require identification of a putative bioactive conformation derived from either experimental evidence, molecular modeling, or conjecture. This requirement may introduce uncertainties into the resulting QSAR models, especially when dealing with structurally diverse data sets containing highly flexible molecules.[19]

Hologram QSAR (HQSAR), recently introduced by Tripos, Inc.,[20] is a novel QSAR method that eliminates the need for determination of 3D structure, putative binding conformations, and molecular alignment. In HQSAR, each molecule in the data set is divided into structural fragments that are then counted in the bins of a fixed length array to form a molecular hologram. This process is similar to the generation of molecular fingerprints employed in database searches[21] and molecular diversity[22] calculations. The bin occupancies of the molecular hologram are structural descriptors (independent variables) encoding compositional and topological molecular information. A linear regression equation that correlates variation in structural information (as encoded in the hologram for each molecule) with variation in activity data is derived through PLS regression analysis to produce a QSAR model. Unlike other fragment-based methods,[23] HQSAR encodes all possible molecular fragments (linear, branched, and overlapping). Optionally, additional 3D information, such as hybridization and chirality, may be encoded in the molecular holograms. Molecular holograms are generated in the same manner as hashed fingerprints where different unique fragments may populate the same holographic bin, allowing the use of a fixed length hologram fingerprint. This hashing procedure emphasizes the importance of patterns of fragment distribution within the hologram bins, which represents the nature of chemical structures more appropriately.[21]

QSAR studies involve two steps: first, descriptors are generated that encode chemical structural information, second, a statistical regression technique is employed to correlate the structural variation, as encoded in the descriptors, with the variation in biological activity. In the present study, three QSAR methods: CoMFA, CoDESSA,[24] and HQSAR were evaluated using three separate data sets. Data sets 1 and 2 contained the same set of structurally diverse molecules but differed with respect to biological endpoints. Data set 3 was composed of a set of congeners exhibiting several degrees of conformational flexibility. All three QSAR methods derive a regression model from PLS analysis; consequently, they differ primarily in the nature of their chemical descriptors. Specifically, CoMFA employs steric and electrostatic field descriptors that encode detailed information concerning intermolecular interaction in three dimensions. CODESSA calculates molecular descriptors on the basis of two-dimensional (2D) and 3D structures and quantum-chemical properties. HQSAR calculates exclusively fragment-based molecular descriptors that are explained in greater detail in the Methodology Section.

By virtue of the differences in chemical descriptors, each of the three QSAR methods will relate molecular structure and properties to estrogenicity in a different way. The specific objective of the present study is to compare CoMFA, CODESSA, and HQSAR as QSAR methods for predicting the binding affinity of a subset of potential EDCs to the ERs. This objective is pursuant to our long-term goal of identifying a QSAR method that can be applied for the rapid screening of large numbers of potential EDCs.

METHODOLOGY

Data Sets for Analysis. The biological activity data used in this study are the relative binding affinity (RBA) obtained from an ER competitive binding assay with labeled endogenous estrogen, 17β-estradiol (E_2). The RBA is 100 times the ratio of the molar concentrations of E_2 and the competing chemical required to decrease the receptor bound radioactivity by 50%.

Data sets 1 and 2 contained the same 31 structurally diverse molecules (Figures 1−5) comprising 19 steroids, four triphenylethylenes, three diethylstilbestrol derivatives, two bis(4-hydroxylphenyl)alkanes, and three phytoestrogens. The RBA values for data sets 1 and 2 were obtained using human ER-α and rat ER-β, respectively.[25] These compounds were used to develop the CoMFA models[10] compared in this paper.

Forty-seven of the compounds contained in data set 3 were largely congeners of the 2-phenylindole prototype structure (Figures 6 and 7).[26−28] Data set 3 also included six structurally diverse estrogens: E_2, ICI 164,384, ICI 182,780, tamoxifen, 4-hydroxytamoxifen, and hexestrol (Figures 1−3). The RBA values for compounds in data set 3 were obtained with calf ER. These data were used to derive the CoMFA and CODESSA models[9] for comparison with the HQSAR models in this paper.

QSAR Methods. All molecular modeling and statistical analyses were performed using Sybyl 6.3[15] and Pirouette 2.03.[29] Procedures for selecting the putative bioactive molecular conformation required for CoMFA and CODESSA, together with rules for structural alignment employed in CoMFA, are described in previous reports.[9,10]

Calculation of CoMFA Descriptors. The aligned molecules were placed in a 3D cubic lattice with 2 Å spacing. Steric (van der Waals) and electrostatic (Coulombic) field descriptors were calculated for each molecule at all lattice points using an sp^3 carbon probe with $+1.0$ charge. Calculated steric and electrostatic energies >30 kcal/mol were truncated to this value. The CoMFA field descriptors were scaled using the CoMFA standard scaling methods[30] provided in Sybyl 6.3.

Calculation of Classical QSAR Descriptors. The CODESSA program was used to generate values for >200 physicochemical descriptors.[24] These descriptors are generally divided into five categories: constitutional, topological,

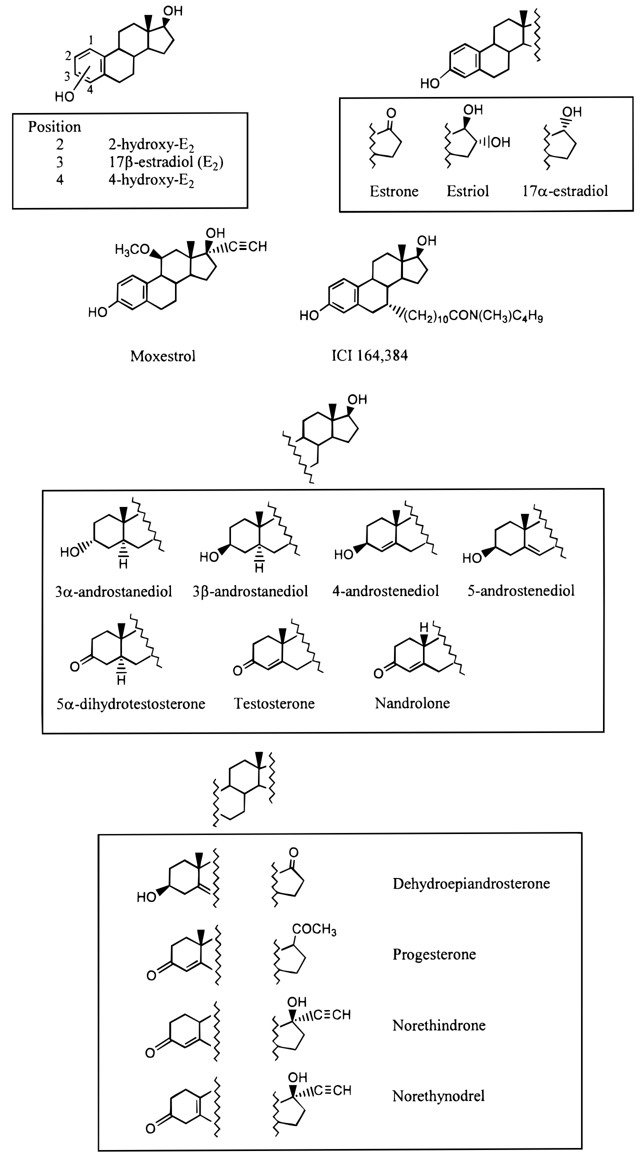


Figure 1. Structures of steroidal compounds in data sets 1 and 2.

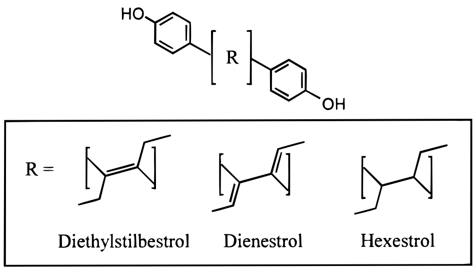
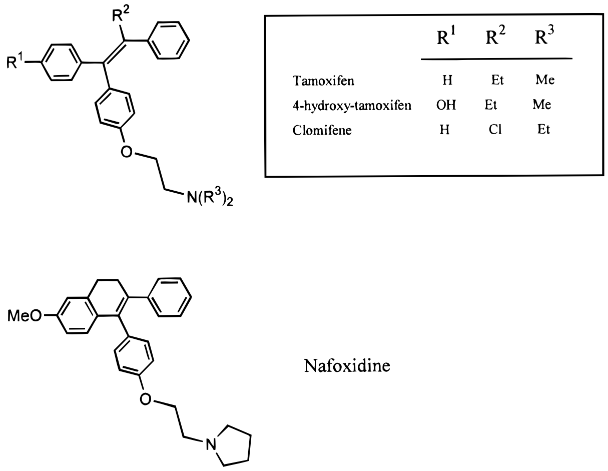


Figure 2. Structures of synthetic estrogens in data sets 1 and 2.



	R^1	R^2	R^3
Tamoxifen	H	Et	Me
4-hydroxy-tamoxifen	OH	Et	Me
Clomifene	H	Cl	Et

Figure 3. Structures of antiestrogens in data sets 1 and 2.

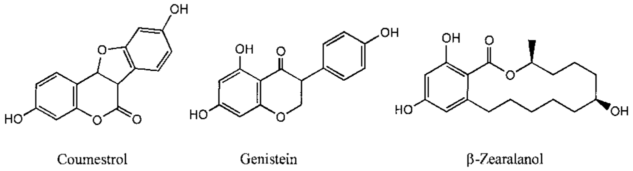


Figure 4. Structures of phytoestrogens in data sets 1 and 2.

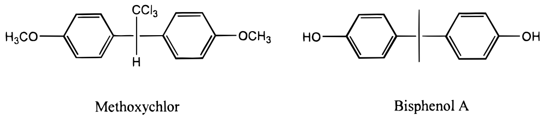
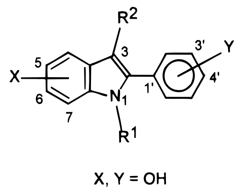


Figure 5. Structures of industrial chemicals in data sets 1 and 2.

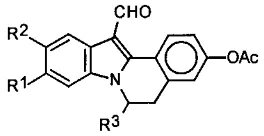
geometrical, electrostatic, and quantum-chemical. The simplest descriptor type is constitutional (e.g., atom counts, molecular weight), which reflects the molecular composition without regard to geometric or electronic structure. Topological descriptors include the Kier and Hall, Randic, and Wiener indices, which are most sensitive to molecular connectivity. Geometrical descriptors, such as moment of inertia and molecular surface area, require the 3D coordinates of the constituent atoms of a molecule. Electrostatic descriptors reflect particular aspects of charge distribution and can be calculated using any of several empirical procedures within the CODESSA program as well as a number of quantum-mechanical approaches. Quantum-chemical descriptors enhance the conventional descriptors by providing information about the internal electronic properties of molecules. CODESSA is capable of computing ~400 descriptors for each molecule. Descriptors for which values are invariant or incalculable for any compound within the data set were excluded from consideration. Of the ~200 remaining descriptors, about half were quantum-chemical in nature. Each set of descriptor values was subjected to autoscaling[31] prior to statistical analysis.

Calculation of HQSAR Descriptors. The following procedure (Figure 8) was used to construct molecular holograms containing the HQSAR descriptors. First, all linear, branched, and overlapping substructural fragments in the size range 4 to 7 atoms were generated for each molecule.[21] All generated fragments from a molecule were then hashed into a fixed length array to produce the molecular hologram. In detail, the procedure is as follows: the SLN (SYBYL Line Notation)[32] for each fragment generated is mapped to a unique integer in the range of 0 to 2^{31} using a CRC (cyclic redundancy check)[33] algorithm. Each integer is then used to select a bin in an integer array of predetermined length (hologram length), the occupancy of which is then incremented by one. The hashing process occurs in cases where the value of the CRC-produced integer is larger than the length of the hologram, and the value of the remainder when the integer is divided by the hologram length is used to identify the array bin whose occupancy was to be incremented. The final array is the molecular hologram, and the bin occupancies are the descriptor variables that encode



Compds	R^1	R^2	Position of X	Y
1a	H	H	6	4'
2a	H	CH_3	6	4'
3a	H	C_2H_5	6	4'
4a	H	H	5	4'
5a	H	CH_3	5	4'
6a	CH_3	H	6	4'
7a	C_2H_5	H	6	4'
8a	C_3H_7	H	6	4'
9a	C_4H_9	H	6	4'
10a	CH_3	CH_3	6	4'
11a	C_2H_5	CH_3	6	4'
12a	C_3H_7	CH_3	6	4'
13a	*i*-C_3H_7	CH_3	6	4'
14a	CH_3	C_2H_5	6	4'
15a	C_2H_5	C_2H_5	6	4'
16a	C_3H_7	C_2H_5	6	4'
17a	CH_3	H	5	4'
18a	C_2H_5	H	5	4'
19a	C_3H_7	H	5	4'
20a	CH_3	CH_3	5	4'
21a	C_2H_5	CH_3	5	4'
22a	C_3H_7	CH_3	5	4'
23a	*i*-C_3H_7	CH_3	5	4'
24a	C_4H_9	CH_3	5	4'
25a	C_5H_{11}	CH_3	5	4'
26a	C_2H_5	C_2H_5	5	4'
27a	C_3H_7	C_3H_7	5	4'
28a	C_2H_5	CH_3	7	4'
29a	C_2H_5	H	6	3'
30a	CH_3	CH_3	6	3'
31a	C_2H_5	CH_3	6	3'
32a	C_3H_7	CH_3	6	3'
33a	C_2H_5	H	5	3'
34a	CH_3	CH_3	5	3'
35a	C_2H_5	CH_3	5	3'
36a	C_3H_7	CH_3	5	3'
ZK119,010	$(CH_2)_6N(CH_2CH_2)_2$	CH_3	5	4'

Figure 6. Structures of 2-phenylindoles in data set 3.



Compds	R^1	R^2	R^3
1b	$OCOCH_3$	H	CH_3
2b	H	$OCOCH_3$	CH_3
3b	$OCOCH_3$	H	C_2H_5
4b	H	$OCOCH_3$	C_2H_5
5b	$OCOCH_3$	H	C_3H_7
6b	H	$OCOCH_3$	C_3H_7
7b	$OCOCH_3$	H	C_4H_9
8b	H	$OCOCH_3$	C_4H_9
9b	$OCOCH_3$	H	C_5H_{11}
10b	H	$OCOCH_3$	C_5H_{11}

Figure 7. Structures of 5,6-dihydroindolo[2,1-α]isoquinolines in data set 3.

molecular structural information. The hologram length (number of array bins) defines the dimensionality of the descriptor space.

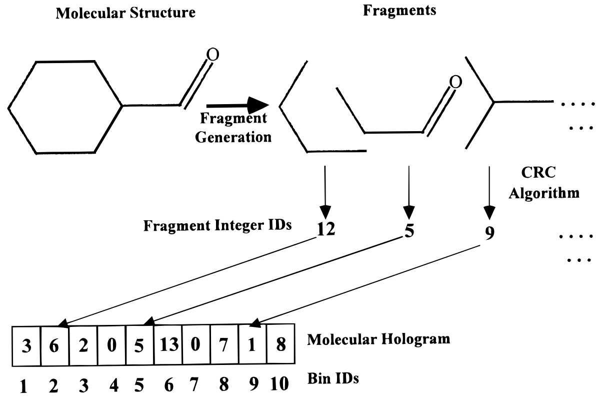


Figure 8. Schematic illustrating generation of a molecular hologram: A molecule is broken into a number of structural fragments that are assigned fragment integer identifications (IDs) using the CRC algorithm. Then each fragment is placed in a particular bin based on its fragment integer ID corresponding to the bin ID. The bin occupancy numbers are HQSAR descriptors that count structural fragments in each bin.

The hashing process greatly reduces the size requirement of a molecular hologram (compared with the case where each unique fragment is counted in its own bin) but leads to a phenomenon called "fragment collision". Identical molecular fragments always generate identical integers through the CRC algorithm and hence will always be counted in the same bin. Typically, because the number of unique fragments contained in a molecule is rather larger than the number of holographic bins, the hashing procedure described will map different integers, and therefore different unique fragments, to the same bin causing fragment collision. In other words, each holographic bin will correspond to several different substructural fragments. Surveying HQSAR models based on a range of different hologram lengths and selecting the hologram length that yields the lowest cross-validated standard error (or highest q^2) minimizes the negative impact of such collisions. The HQSAR module provides 12 default hologram lengths that have been found to yield predictive models on a number of test data sets. These default hologram lengths are prime numbers such that each provides a unique set of fragment collisions.

The exact model produced by HQSAR is dependent not only on the hologram length but also on the information contained in the generated fragments. The particular nature of substructural fragments generated by HQSAR and, consequently, the information contained in the resultant molecular holograms can be altered through the settings of seven parameters. These hologram construction parameters are divided into two classes: *fragment size and fragment distinction.* The two *fragment size* parameters, minimum and maximum fragment size, determine the maximum and minimum number of atoms in any one fragment (the default values for these parameters are 4 and 7, respectively). *Fragment distinction* parameters describe what information from the original molecule is retained in the fragment in terms of *atoms*, *bonds*, *connections*, *hydrogens*, and *chirality.* Table 1 and Figure 9 depict how these different parameter settings affect the information contained in the generated fragments and lead to the generation of distinct fragments from the same portion of the original molecule.

PLS-QSAR. Predictive QSAR models were produced using separate PLS analyses of the three data sets to correlate variation in biological activity with variation in the descriptors described in the previous sections. The optimum number of principal components (PCs) corresponding to the smallest standard error of prediction was determined by the Leave-

Table 1. Definition of Fragment Parameters in HQSAR

parameter	definition
atoms	The *atoms* parameter enables fragments to be distinguished based on elemental atom types; for example, allowing NH_3 to be distinguished from PH_3.
bonds	The *bonds* parameter enables fragments to be distinguished based on bond orders; for example, in the absence of hydrogen, allowing butane to be distinguished from 2-butene.
connections	The *connections* parameter provides a measure of atomic hybridization states within fragments; that is, *connections* causes HQSAR to keep track of how many connections are made to constituent atoms and the bond order of those connections.
hydrogens	By default, HQSAR ignores the hydrogen atoms during fragment generation. The *hydrogens* parameter overrides this behavior.
chirality	The *chirality* parameter enables fragments to be distinguished based on atomic and bond stereochemistry. Thus, stereochemistry allows *cis* double bonds to be distinguished from their *trans* counterparts, and *R*-enantiomers to be distinguished from *S*-enantiomers at all chiral centers.

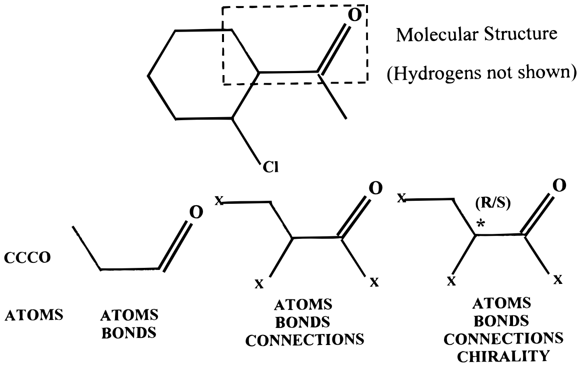


Figure 9. Schematic illustrating different fragment parameters in HQSAR.

Table 2. Summary of the Key Statistical Parameters Obtained for Each QSAR Model

datasets	statistics	CoMFA	HQSAR	CODESSA
1	q^2	0.70	0.67	0.46
	r^2	0.95	0.88	0.79
	PCs	4	4	2
2	q^2	0.60	0.68	0.61
	r^2	0.95	0.91	0.92
	PCs	4	5	4
3	q^2	0.61	0.53	0.54
	r^2	0.97	0.93	0.68
	PCs	9	9	3

One-Out (LOO) cross-validation procedure. By this procedure, each compound is systematically excluded once from the data set, after which its activity is predicted by a model derived from the remaining compounds. The predicted activities of the "left out" compounds allow the calculation of q^2 and cross-validated standard error. Using the optimal number of PCs, the final PLS analysis was carried out without cross-validation to generate a predictive QSAR model with a conventional correlation coeffficient r^2 and a non-cross-validated standard error.

RESULTS

Quality of the QSAR Models. The quality of a QSAR model can be assessed in terms of various statistical measures. The values of r^2 and q^2 are normally accepted as statistical measures of merit for a QSAR model. In many QSAR studies, the criterion $r^2 \geq 0.9$ is employed to decide whether a model is internally self-consistent. It should be noted that r^2 makes no assessment of the intrinsic precision or accuracy of the data itself. The value of q^2, derived from the LOO cross-validation procedure, tests the stability of the model through perturbation of the regression coefficients by consecutively omitting each compound during the model generation procedure. Consequently, q^2 can be considered a measure of the ability of the model to interpolate within the training set population. The criterion $q^2 \geq 0.5$ is normally adopted in many CoMFA studies for determining the acceptability of the model for this purpose.[34] Values of the r^2 and q^2 associated with the three QSAR models for each of three data sets are given in Table 2. In this example, only 2D connectivity, bond and elemental atom-type information *(atoms* and *bonds* parameters turned on) was used in the HQSAR calculations. It is seen that all three QSAR models exceeded the $q^2 \geq 0.5$ criterion. In terms of goodness of fit, CoMFA models provided the highest r^2 values accounting for at least 95% of the variation in biological activity in all data sets. Individual r^2 values for each data set were slightly lower for HQSAR than for CoMFA models. Importantly, the average r^2 value for all three data sets exceeded 0.90 for both HQSAR and CoMFA. CODESSA yielded a good QSAR model for data set 2, but lower r^2 values for data sets 1 and 3 thus indicating that further work on CODESSA may be required.

Although r^2 and q^2 are important for validating the quality of a QSAR model, these parameters alone fail to account for other factors. One such factor is the number of principal components (degrees of freedom) that should be considered when comparing different QSAR models derived from an individual data set. The value of r^2 generally increases as more PCs are included in the model. Thus, it would seem reasonable to scale a statistical parameter of choice by the number of PCs. Indeed, the primary motivation for using the PLS method is to build the most predictive model that fits the known biological data (high q^2 and r^2, respectively) with the fewest number of PCs to avoid overfitting of data points. Another factor is the range of biological activity within the data set, which also should be considered during the comparison of the quality of QSAR models across different data sets. Given two QSAR models that have the same r^2 (or q^2) value, the model derived from the data set with the larger biological activity range is more valid than that obtained with the smaller activity range.

Alternatively, the standard error and cross-validated standard error can be used as measures of goodness of fit and predictivity. Although several ways exist to calculate the standard error for a regression equation, the number of degrees of freedom should be factored in when comparing different models. A more effective measure of model goodness of fit is the ratio of the standard error to the activity range. One advantage of explicitly including the range of biological activity is that the performance of separate QSAR models can be compared across different data sets. This ratio should generally be <10% for good QSAR models.[35] The

Table 3. Ratio of the Standard Error to the Activity Range, Given as a Percentage

dataset	PLS	CoMFA	HQSAR	CODESSA
1	cross-validated	15.7	16.4	20.3
	non-cross-validated	6.3	9.9	12.6
2	cross-validated	17.4	15.9	17.1
	non-cross-validated	6.5	8.5	7.6
3	cross-validated	15.0	16.5	16.5
	non-cross-validated	4.5	6.3	13.8

Table 4. Observed Versus Predicted Log RBA Values[a]

dataset	cpd	observed	CoMFA	CODESSA	HQSAR
1	**1**	<−2.0	-1.95	−3.84	−2.48
	2	<−2.0	-2.10	−3.97	−2.48
	3	<−2.0	-2.22	−3.31	−2.59
	4	<−3.0	−2.41	−0.61	−2.36
2	**1**	<−2.0	−2.11	−3.78	−1.81
	2	<−2.0	−2.32	−4.57	−1.81
	3	<−2.0	−2.50	−2.69	−2.48
	4	<−3.0	−2.05	−0.50	−3.03

[a] Obtained by the three QSAR methods under study for the following test compounds in datasets 1 and 2: 5α-androstanedione (**1**), 5*β*-androstanedione (**2**), 4-androstenedione (**3**), and corticosterone (**4**).

Table 5. Summary of Steps in Developing QSAR Models for CoMFA, HQSAR, and CODESSA

step	CoMFA	HQSAR	CODESSA
(1) determine conformation	required	not required	required
(2) generate descriptors[a]	determine alignment	generate hologram	AMPAC calculation
(3) statistics (LOO/PLS)	descriptor space (>2000)	descriptor space (<500)	descriptor space (<500)

[a] In each case, only the "rate-determining" step is listed.

percentage ratios of the standard error to the activity range of QSAR models in this study for both the cross-validated and non-cross-validated PLS analyses are summarized in Table 3. The values of this ratio for both CoMFA and HQSAR models are low, further substantiating their statistical validity. In contrast, higher values of this ratio are seen for the CODESSA models in data sets 1 and 3. This observation is consistent with the large standard error associated with the CODESSA models and with the low value of r^2 noted in Table 2.

Predictions for Test Compounds. Four compounds from data sets 1 and 2, namely 5α-androstanedione (**1**), 5*β*-androstanedione (**2**), 4-androstenedione (**3**), and corticosterone (**4**), were excluded from the training set to serve as test compounds to evaluate the predictive ability of the present QSAR models. These particular compounds were selected, in part, because their biological data were reported as "less than" values. Although the approximate nature of the RBA values for these compounds precluded their use in the training sets, these RBA values could still be compared with those predicted by each of the three QSAR models. The results for the test-set compounds are summarized in Table 4, in which the observed log RBA values are listed along with the corresponding log RBA values predicted by the three QSAR models based on both the human ER-α (data set 1) and rat ER-*β* data (data set 2). The log RBA values predicted by CoMFA and HQSAR are highly consistent with the experimental data and with each other. For **1**, **2**, and **3**, CoMFA and HQSAR correctly predicted that the log RBA values are indeed <−2.0 or close to −2.0. Although only HQSAR correctly predicted that the log RBA value for **4** in data set 2 is <−3.0, the log RBA values predicted by CoMFA and HQSAR are in agreement with each other and are in reasonable agreement with the experimentally determined limit. The corresponding log RBA values predicted by the CODESSA model appear less satisfactory. Notably the CODESSA-predicted activities of **4**, for both biological endpoints, were in poor agreement with experiment, being >2 log units from the maximum experimentally determined limit. Additionally, although the CODESSA-predicted values for **1**, **2**, and **3** were in accordance with the experimental limitations, they were in poor agreement with the CoMFA and HQSAR predicted values. Finally, CODESSA predicted activities for **1** and **2** were outside the range of activities found in the training datasets used to produce the QSAR models.

Utility of the QSAR Approaches for Screening. QSAR screening of a large number of chemicals for endocrine disruption potential requires a highly practical and accurate QSAR method. Three criteria for practicality were included in this study; namely, computation time, reproducibility, and convenience. Computation time is a significant concern when screening huge numbers of compounds. Reproducible QSAR models provide an opportunity for different investigators to compare and validate prediction results. A convenient QSAR method allows a non-expert user to make biological activity predictions more readily. A fast, reproducible, user-friendly QSAR prediction method offers major advantages for the routine screening of large chemical databases for potential EDCs.

The key molecular modeling and statistical analysis processes required for CoMFA, HQSAR, and CODESSA model development are listed in Table 5. The first step in both CoMFA and CODESSA studies is the determination of the putative ligand binding conformation. Because experimental evidence about ligand−receptor binding conformations is usually lacking, the bioactive conformation must be postulated based on information about the receptor binding site and/or the common conformational space accessible to different known ligands. In lieu of such information, the global minimum-energy conformation is commonly selected. Regardless of the choice, a considerable amount of time and expertise is required for molecular modeling. In contrast to CoMFA and CODESSA, HQSAR requires only information about the 2D molecular structure, requiring little or no molecular modeling. Generation of descriptors using CoMFA and CODESSA involve time-consuming processes that can be carried out effectively only by expert modelers; for examples, structural alignment in CoMFA and semi-empirical quantum mechanical (AMPAC/MOPAC) calculations in CODESSA. In contrast, generation of molecular holograms as the chemical descriptors in HQSAR takes considerably less time and expertise. It is worthwhile to mention that the construction of the regression equation through standard PLS analysis takes less time in HQSAR than in CoMFA inasmuch as the number of descriptors generated is generally far less.

Due to the dependence of CODESSA and CoMFA models on molecular conformation and (CoMFA) structural alignment, in which small perturbations can become magnified

Table 6. Influence of Various Fragment-Type Parameters on the r^2 and q^2 of the Resulting HQSAR Model

		additonal option added[a]				
data set	statistics	none[a]	Con[b]	H[c]	Con-H	Chi[d]
1	q^2	0.67	0.68	0.51	0.54	0.67
	r^2	0.88	0.88	0.81	0.90	0.91
2	q^2	0.68	0.65	0.40	0.42	0.64
	r^2	0.91	0.88	0.57	0.63	0.88
3	q^2	0.53	0.68	0.59	0.54	0.61
	r^2	0.93	0.96	0.94	0.87	0.93

[a] In every case, the *Atoms* and *Bonds* flags are turned on. [b] Con, connectivity flag is on. [c] H, hydrogens flag is on. [d] Chi, chirality option is used by combining with Con-H.

in the final QSAR model, much care must be taken when generating these models to ensure reproducibility. Because the calculation of HQSAR descriptors from counts of substructural molecular fragments is straightforward, model reproducibility is readily achieved in minimal time.

Evaluation of the Fragment Parameters in HQSAR. Based on our initial results, which demonstrated the utility of HQSAR for screening large databases, the technique was investigated more thoroughly by varying the fragment type and length parameters. The data in Table 6 shows that predictive HQSAR models are readily derived using only elemental and bond-type information. Incorporating hydrogen-containing fragments into molecular holograms (turning on the *Hydrogens* parameter) appears to decrease the signal-to-noise ratio in molecular holograms. Thus, PLS has greater difficulty in determining a predictive model that fits the known activity data, as evidenced by the lower q^2 and r^2 values in data sets 1 and 2. The inclusion of atomic hybridization and chirality information also failed to improve significantly the quality of the HQSAR models.

Fragment size parameters control the minimum and maximum length of fragments to be included in the hologram fingerprint. As mentioned previously, molecular holograms are formed by the generation of all linear, branched, overlapping fragments between $\boldsymbol{M}$ and $\boldsymbol{N}$ atoms in size. The parameters $\boldsymbol{M}$ and $\boldsymbol{N}$ can be changed to include smaller or larger fragments in the holograms. Default fragment lengths of $\boldsymbol{M} = 4$ and $\boldsymbol{N} = 7$ are provided. The HQSAR results for six different fragment sizes for data set 1 are summarized in Table 7. The highest values for both r^2 and q^2 were obtained for fragment lengths of 6−9; however, neither r^2 or q^2 showed much sensitivity to fragment length. Overall, minor alteration of any of the HQSAR parameter settings from those provided as default failed to alter the quality of generated QSAR models to any significant extent.

DISCUSSION

Three QSAR models were developed for each of three data sets. These models were compared using several statistical measures, including r^2, q^2, and the ratio of standard error to the activity range. A number of variations on the basic HQSAR model (which had only the *atoms*, *bonds* parameters turned on and used default fragment lengths) were also developed, using information on *connectivity*, *hydrogens*, and *chirality*. The variants, which increase hologram information content, did not provide any general improvement in the basic model as measured by r^2 and q^2.

Table 7. Comparison of Different Fragment Lengths on the Resulting HQSAR Models[a] for Data Set 1

fragment length	q^2	r^2
2−5	0.67	0.86
3−6	0.60	0.81
4−7	0.67	0.88
5−8	0.70	0.92
6−9	0.73	0.94
7−10	0.68	0.91

[a] Fragment type is only *Atoms* and *Bonds*.

Data set 2 describes ligand binding to the ER-β,[24] a recently discovered ER different from but with homology to the classical ER, now termed ER-α (data set 1). The present study, which includes HQSAR and CODESSA models, compliments and extends our early development of QSAR models for these ERs using CoMFA.[10]

In the present application, CoMFA yielded the best QSAR models in terms of self-consistency and ability to interpolate within the training set population. Because the molecular descriptors in CoMFA encode for molecular shape and charge distribution in 3D space, it is not surprising that CoMFA was best able to capture the salient features associated with molecular recognition in ER binding. Furthermore, information derived from CoMFA models can be visualized and employed to determine the 3D properties of the molecules under study that may be responsible for activity at the ER. Although the HQSAR models under comparison included only elemental and bond-type information, the quality of the HQSAR models was comparable with those from CoMFA. Elemental and bond-type information included in molecular holograms is compositional and topological in nature.

Similar information is also included among the CODESSA descriptors. However, HQSAR and CODESSA differ fundamentally in the way they encode the topological features of molecules. Typical topological descriptors in CODESSA, such as the Kier−Hall and Randic−Wiener indices, compress molecular topological information into a single value. This reduction of connectivity information to a single number leads to a degree of information loss. In contrast, topological information in HQSAR is encoded in structural fragments that are distributed into molecular holograms for selection and processing by PLS. This process leads to a lesser degree of topological information loss. Differences in topological features for a set of molecules are well represented in the HQSAR descriptors. In contrast to CoMFA and HQSAR, CODESSA is apparently lacking a sufficient number of descriptors that are readily selected by PLS to correlate with estrogenic activity (with specific reference to data sets 1 and 3). The resolution of these issues associated with CODESSA, although not an objective of the present study, may emerge from the selection and optimization of the collection of descriptors by using variable selection methods such as genetic algorithms in conjunction with PLS[36] or other statistical methods[37] such as artificial neural networks.[38]

Applications of QSAR methods continue to grow. One general application is to identify important structural features relating a specific biological activity for lead discovery and/or optimization in drug design. CoMFA and classical QSAR are more suitable for this purpose. A second application is mass screening of large chemical databases to predict specific biological activities. In the present case, new legislation requires the screening of >80 000 chemicals for potential endocrine disrupting activity. A major category includes estrogenic chemicals that act via binding to both the ER-α and -β subtypes. Because inactive chemicals can be effectively separated from active molecules based on 2D descriptors using hierarchical clustering methods,[39] the challenge is to develop QSAR procedures that identify active chemicals with a high degree of confidence. Additionally, combinatorial chemistry techniques are dramatically increasing the number of chemicals under consideration for product development. Therefore, it is important to have a QSAR technique that offers not only consistent and reproducible predictivity, but also a fast and convenient procedure. HQSAR models appear well suited for such applications. Because the CoMFA and CODESSA requirements for 3D structure, bioactive conformation, and molecular alignment are eliminated in HQSAR, the HQSAR method provides shorter computation time, simple reproducibility, and convenience. These three factors combined with the ability to generate a robust model give the HQSAR technique significant advantages for use in screening large datasets of chemicals (e.g., EDCs). The marginally better statistical results associated with the CoMFA-generated models do not compensate for these practical limitations. Current CODESSA models also require one to know or postulate the bioactive conformation and may include time-consuming quantum-mechanical calculations. Additionally, CODESSA models perform less satisfactorily than either CoMFA or HQSAR models for the present datasets, according to statistical measurements.

CONCLUSION

Three different techniques for the generation of QSAR models—CoMFA, CODESSA, and HQSAR—were evaluated for their utility (predictive, fast, readily reproducible) to screen large numbers of compounds for estrogenic activity. The CoMFA models emerging from this study were of good-to-excellent quality (high r^2) and exhibited good predictive ability for interpolation within the training set population (high q^2). Predictions made with CoMFA models on four compounds excluded from the training set were in good agreement with the experimentally determined values. Moreover, information derived from CoMFA models can be employed to identify specific molecular factors responsible for the differing activities in a group of molecules. Although CoMFA models are of high quality and can give indication of structural differences responsible for differing biological activities, they can be time consuming to construct because they require determination of suitable molecular conformations and a structural alignment of the molecules under study.

For the present three data sets under investigation, the QSAR models generated based on CODESSA descriptors with implementation of PLS have relatively lower quality. Further analysis and validation of this technique is suggested before it can be used as a prioritizing method for potential EDCs.

QSAR models generated through the HQSAR technique have comparable quality to those of CoMFA. The HQSAR method also showed good agreement with both experiment and CoMFA in the prediction of the four compounds excluded from the training datasets. Furthermore, because HQSAR employs counts of substructural molecular fragments as descriptors and requires no 3D structures or molecular alignment, it is both fast and reproducible.

Because of new legislation, the EPA has been mandated to develop a screening and testing program for potential endocrine disrupting chemicals. QSAR methodologies would be useful as a prioritization tool for the large number of compounds requiring testing before the use of *in vitro* and *in vivo* assays. Such an approach requires the QSAR technique employed to possess certain fundamental qualities: good predictivity, speed, and ease of use. Among the QSAR methods examined, HQSAR appears to offer many attractive features that portend its utility for prioritizing potential EDCs for subsequent toxicological testing and risk assessment.

REFERENCES AND NOTES

(1) Kavlock, R. J.; Daston, G. P.; DeRosa, C.; Fenner-Crisp, P.; Gray, L. E.; Kaattari, S.; Lucier, G.; Luster, M.; Mac, M. J.; Maczka, C.; Miller, R.; Moore, J.; Rolland, R.; Scott, G.; Sheehan, D. M.; Sinks, T.; Tilson, H. A. Research needs for the risk assessment of health and environmental effects of endocrine disrupters: A report of the U.S. EPA-sponsored workshop. *Environ. Health Perspect.* **1996**, *104*, 715−740.

(2) *Compilation of Laws Enforced by the U.S. Food and Drug Administration and Related Statutes*; Vol. 2; U.S. Government Printing Office: Washington, D.C., 1996.

(3) *Safe Drinking Water Act Amendment of 1996*, Public Law 104−182, 104th Congress, 1996.

(4) Patlak, M. A testing deadline for endocrine disrupters. *Environ. Sci. Technol.* **1996**, *30*, 540A−544A.

(5) Warr, W. Combinatorial chemistry and molecular diversity. An Overview. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 134−140.

(6) Broach, J. R.; Thorner, J. High throughput screening for drug discovery. *Nature* **1996**, *384(Supp)*, 14−16.

(7) Katzenellenbogen, J. A. The structural pervasiveness of estrogenic activity. *Environ. Health Perspect.* **1995**, *103*, 99−101.

(8) Anstead, G. M.; Carlson, K. E.; Katzenellenbogen, J. A. The estradiol pharmacophore: ligand structure-estrogen receptor binding affinity relationships and a model for the receptor binding site. *Steroids* **1997**, *62*, 268−303.

(9) Tong, W.; Perkins, R.; Strelitz, R.; Collantes, E. R.; Keenan, S.; Welsh, W. J.; Branham, W. S; Sheehan, D. M. Quantitative structure-activity relationships (QSARs) for estrogen binding to estrogen receptor: Predictions across species. *Environ. Health Perspect.* **1997**, *105(10)*, 1116−1124.

(10) Tong, W.; Perkins, R.; Xing, L.; Welsh, W. J.; Sheehan, D. M. QSAR models for binding of estrogenic compounds to estrogen receptor α and β subtypes. *Endocrinology* **1997**, *138*, 4022−4025.

(11) Waller, C. L.; Minor, D. L.; Mckinney, J. D. Examination of the estrogen-receptor binding affinities of polychlorinated hydroxybiphenyls using three-dimensional quantitative structure-activity relationships. *Environ. Health Perspect*. **1995**, *103*, 702−707.

(12) Bradbury, S. P.; Mekenyan, O. G.; Ankley, G. T. Quantitative structure-activity relationships for polychlorinated hydroxybiphenyl estrogen receptor binding affinity - An assessment of conformer flexibility. *Environ. Tox. Chem.* **1996**, *15*, 1945−1954.

(13) Gantchev, T. G.; Ali, H.; van Lier, J. E. Quantitative structure-activity relationships/comparative molecular field analysis (QSARs/CoMFA) for receptor-binding properties of halogenated estradiol derivatives. *J. Med. Chem.* **1994**, *37*, 4164−4176.

(14) Waller, C. L.; Oprea, T. I.; Chae, K.; Park, H. K.; Korach, K. S.; Laws, S. C.; Wiese, T. E.; Kelce, W. R.; Gray, L. E. Ligand-based identification of environmental estrogens. *Chem. Res. Toxicol.* **1996**, *9*, 1240−1248.

(15) Cramer, R., III; Patterson, D. E.; Bunce, J. D. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959−5967.

(16) Collantes, E.; Tong, W.; Welsh, W. J. Use of moment of inertia in comparative molecular field analysis to model chromatographic retention of nonpolar solutes. *Anal. Chem.* **1996**, *68*, 2038−2043.

(17) Tong, W.; Collantes, E. R.; Welsh, W. J.; Berglund, B.; Howlett, A. Derivation of a pharmacophore model for anandamide using constrained conformational searching and comparative molecular field analysis (CoMFA). *J. Med. Chem.*, in press.

(18) Welsh, W. J.; Tong, W.; Collantes, E. R. Heats of sublimition and formation of polycyclic aromatic hydrocarbons (PAHs) derived from comparative molecular field analysis (CoMFA): Application of moment of inertia for molecular alignment. *Thermochim. Acta* **1996**, *290*, 55−64.

(19) Tong, W.; Collantes, E. R.; Chen, Y.; Welsh, W. J. A comparative molecular field analysis study of *N*-benzylpiperidines as acetylcholinesterase inhibitors. *J. Med. Chem.* **1996**, *39*, 380−387.

(20) HQSAR is a product of Tripos, Inc., St. Louis, MO 63144.

(21) James, C. A.; Weininger, D. *Daylight Theory Manual*; Daylight Chemical Information Systems, Inc.: 27401 Los Altos, Suite #370, Mission Viejo, CA 92691.

(22) Turner, D. B.; Tyrrell, S. M; Willett, P. Rapid quantification of molecular diversity for selective database acquisition. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 18−22.

(23) Rosenkranz, H. S.; Cunningham, A.; Klopman, G. Identification of a 2-D geometric descriptor associated with non-genotoxic carcinogens and some estrogens and antiestrogens. *Mutagenesis* **1996**, *11*, 95−100.

(24) CODESSA is a product of Semichem, 7128 Summit, Shawnee, KS 66216.

(25) Kuiper, G. G. J. M.; Carlsson, B.; Grandien, K.; Enmark, E.; Haggblad, J.; Nilsson, S.; Gustafsson, J-Å. Comparison of the ligand binding specificity and transcript tissue distribution of estrogen receptors α and *β*. *Endocrinology* **1997**, *138*, 863−870.

(26) von Angerer, E.; Prekajac, J.; Strohmeier, J. 2-Phenylindoles. Relationship between structure, estrogen receptor affinity, and mammary tumor inhibiting activity in the rat. *J. Med. Chem.* **1984**, *27*, 1439−1447.

(27) Polossek, T.; Ambros, R.; von Angerer, S.; Brandl, G.; Mannschreck, A.; van Angerer, E. 6-Alkyl-12-formylindolo[2,1-*a*]isoquinolines. Synthesis, estrogen receptor binding affinities, and stereospecific cytostatic activity. *J. Med. Chem.* **1992**, *35*, 3537−3547.

(28) von Angerer, E.; Biberger, C.; Holler, E.; Koop, R.; Leichtl, S. 1-Carbamoylalkyl-2-phenylindoles: relationship between side chain structure and estrogen antagonism. *J. Steroid Biochem. Molec. Biol.* **1994**, *49*, 51−62.

(29) InfoMetrix, Inc., P. O. Box 1528, Woodinville, Washington 98027.

(30) Cramer, R. D., III Partial least square (PLS): Its strengths and limitations. *Perspect. Drug Dis. Design* **1993**, *1*, 269−278.

(31) Livingstone, D. *Data analysis for chemists—applications to QSAR and chemical product design*; Oxford University. Oxford, New York, 1995.

(32) Ash, S.; Cline, M.; Homer, R. W.; Hurst, T.; Smith, G. B. SYBYL Line Notation (SLN): A versatile language for chemical structure representation. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 71−79.

(33) Knuth, D. E. *Sorting and searching*; Addison-Wesley: Reading, MA.

(34) Cramer, R. D., III; Bunce, J. D.; Patterson, D. E. Crossvalidation, bootstrapping, and partial least squares compared with multiple regression in conventional QSAR studies. *Quant. Struct.−Act. Relat.* **1988**, *7*, 18−25.

(35) Apex-3D 95.0 User Guide, Biosym/MSI, 9685 Scranton Road, San Diego, CA 92121.

(36) Hasegawa, K.; Miyashita, Y.; Funatsu, K. GA strategy for variable selection in QSAR studies: GA-based PLS analysis of calcium channel antagonists. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 306−310.

(37) Kubinyi, H. Evolutionary variable selection in regression and PLS analysis. *J. Chemometr.* **1996**, *10*, 119−133.

(38) So, S-S; Karplus, M. Evolutionary optimization in quantitative structure-activity relationship: an application of genetic neural networks. *J. Med. Chem.* **1996**, *39*, 1521−1530.

(39) Brown, R. D.; Martin, Y. C. Use of structure-activity data to compare structure-based clustering methods and descriptors for use in compound selection. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 572−584.

CI980008G